DATAFLAIR WEB SERVICES PVT LTD

Weblog Analysis

Analyze Weblogs using MR

Anish P

This Document lists Weblogs Analysis using Map-Reduce. The document also contains links for the data which can be downloaded for POC purpose.

Contents

Weblog Analysis:

Weblogs

What is Weblogs

Weblogs are where web server (like apache) records events like visitors to your site and problems it's encountered. Your web server records all the visitors to your site. There you can see what files users are accessing, how the web server responded to requests, and other information like what kind of web browsers visitors are using, etc.

Sample Weblog Data

```
133.128.48.53 - - [01/Jan/2012:01:55:42 +0530] "GET /mobiles/smart-phones/sony-xperia-m-dual-android-smart-phone-white.html HTTP/1.1" 200 1466 "Mozilla/5.0 (Windows NT 6.2) AppleWebKit/537.17 (KHTML, like Gecko) Chrome/24.0.1312.56 Safari/537.17" "-"
```

Format of Weblogs

1. Remote-IP
2. Remote-log-name
3. user
4. time
5. request-string
6. status-code
7. byte-string
8. user-agent
9. referral

Data for POC

- https://www.dropbox.com/s/n7u05t9borh084f/weblogs_1_thousand_rec.txt?dl=0
- https://www.dropbox.com/s/x3ao2h24jk3u5tz/weblogs_10_thousand_rec.txt?dl=0
- https://www.dropbox.com/s/mfsypcfyjyo53se/weblogs_1_lakh_rec.txt?dl=0
- https://www.dropbox.com/s/w3akmb8531xe94z/weblogs_10_lakh_rec.txt?dl=0

KPI – 1: Parse weblogs into structured format

Input (Raw Log Record)

```
50.57.190.149 - - [22/Apr/2012:07:12:41 +0530] "GET /computers/laptops.html?brand=819 HTTP/1.0"
200 12530 "-" "-"
```

Output (Processed Log Record)

```
50.57.190.149	-	-	22/Apr/2012:07:12:41 +0530	GET
/computers/laptops.html?brand=819 HTTP/1.0	computers	-	-	laptops.html
	brand=819	200	12530	-	-
```

Format of Input Data

1. remote-IP
2. remote-log-name
3. user
4. time
5. request-string
6. status-code
7. byte-string
8. user-agent
9. referral

Tasks

1. Fields should be ‘Tab’ separated
2. Remove ‘[ ]’ from date-time field
3. Remove ‘””’ from request-string
4. Parse request-string into structured format /cat-1/cat-2/cat-3/cat-4/page?param → cat-1 cat-2 cat-3 cat-4 page param
5. Remove ‘””’ from user-agent
6. Remove ‘””’ from referral
7. Handle bad records, store bad records in a specific file

KPI – 2: Count of page views by individual user

Input (Processed Log Record)

```
50.57.190.149   -           -           22/Apr/2012:07:12:41 +0530     GET
/computers/laptops.html?brand=819 HTTP/1.0  computers       -           -           laptops.html
            brand=819        200        12530   -           -
```

Format of Processed weblogs (Input Data)

1. remoteIP
2. remotelogname
3. user
4. time
5. request-string
6. category-1
7. category-2
8. category-3
9. page
10. param
11. status-code
12. byte-string
13. user-agent
14. referral

Task

- Take the pre-processed weblogs as input
- Count the number of page visited by each individual user (User wise page-visit distribution)

KPI – 3: Count of page views by catagery-1

Input (Processed Log Record)

```
50.57.190.149   -           -           22/Apr/2012:07:12:41 +0530     GET
/computers/laptops.html?brand=819 HTTP/1.0  computers       -           -           laptops.html
            brand=819       200       12530   -           -
```

Format of Processed weblogs (Input Data)

1. remoteIP
2. remotelogname
3. user
4. time
5. request-string
6. category-1
7. category-2
8. category-3
9. page
10. param
11. status-code
12. byte-string
13. user-agent
14. referral

Task

- Take the pre-processed weblogs as input
- Count the number of page visited based on categories (Category wise page-visit distribution)